Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences

Christophe Botella, Maximilien Servajean, Pierre Bonnet, Alexis Joly

▶ **To cite this version:**

Christophe Botella, Maximilien Servajean, Pierre Bonnet, Alexis Joly. Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences. CLEF 2019 - Conference and Labs of the Evaluation Forum, Sep 2019, Lugano, Switzerland. hal-02190170

HAL Id: hal-02190170
https://hal.science/hal-02190170

Submitted on 22 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences

Botella Christophe[1,2,3], Servajean Maximilien[5], Bonnet Pierre[3,4], Joly Alexis[1]

[1] INRIA Sophia-Antipolis - ZENITH team, LIRMM - UMR 5506 - CC 477, 161 rue Ada, 34095 Montpellier Cedex 5, France.
[2] INRA, UMR AMAP, F-34398 Montpellier, France.
[3] AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France.
[4] CIRAD, UMR AMAP, F-34398 Montpellier, France.
[5] LIRMM, Université Paul Valéry, University of Montpellier, CNRS, Montpellier, France

Abstract. The GeoLifeCLEF challenge aim to evaluate location-based species recommendation algorithms through open and perennial datasets in a reproducible way. It offers a ground for large-scale geographic species prediction using cross-kingdom occurrences and spatialized environmental data. The main novelty of the 2019 campaign over the previous one is the availability of new occurrence datasets: (i) automatically identified plant occurrences coming from the popular Pl@ntnet platform and (ii) animal occurrences coming from the GBIF platform. This paper presents an overview of the resources and assessment of the GeoLifeCLEF 2019 task, synthesizes the approaches used by the participating groups and analyzes the main evaluation results. We highlight new successful approaches relevant for community modeling like models learning to predict occurrences from many biological groups and methods weighting occurrences based on species infrequency.

Keywords: LifeCLEF, biodiversity, environmental data, species recommendation, evaluation, benchmark, Species Distribution Models, methods comparison, presence-only data, model performance, prediction, predictive power

1 Introduction

The automatic prediction of the species most likely to be observed at a given location is an important issue for many areas such as biodiversity conservation, land management or environmental education. First, it could improve species identification processes and tools by reducing the list of candidate species observable at a given site (whether automated, semi-automatic or based on traditional

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

field guides or flora). More generally, it could facilitate biodiversity inventories and compliance with regulatory obligations for the environmental integration of development projects. Finally, it could be used for educational purposes through biodiversity discovery applications offering functionalities such as contextualized educational pathways.

In the context of LifeCLEF evaluation campaign 2019 [6], the objective of the GeoLifeCLEF challenge is to evaluate the state of the art of species prediction methods over the long term and with a view to reproducibility. To achieve this, the challenge freely provides researchers with large-scale, documented and accessible data sets over the long term. Concretely, the aim of the challenge is to predict the list of species that are the most likely to be observed at a given location. Therefore, we provide a large training set of species occurrences and a set of environmental rasters that characterize the environment in a quantitative and qualitative way at any position in the territory. Indeed, it is usually not possible to learn a species distribution models directly from spatial positions because of the limited number of occurrences and the sampling bias. What is usually done in ecology is to predict the distribution of species based on a representation in environmental space, typically a characteristic vector composed of climatic variables (mean temperature at that location, precipitation, etc.) and other variables such as soil type, land cover, distance to water, etc. GeoLifeCLEF's originality is to encourage the extension of this approach to learning a more complex representation space that takes into account various input data such as environmental descriptors, their spatial structure and the known biotic context. Therefore, we provide tools to facilitate the extraction of environmental tensors that can be easily used as input data to models such as convolutional neural networks.

In 2019, the provided data was significantly enriched and several methodological improvements have been made. In more details, the new features introduced are as follows:

1. Pl@ntNet occurrences: to increase the amount of plant occurrences in the training set, we completed the publicly available data from the GBIF[6] with user-generated observations of the Pl@ntNet mobile application [1]. These data are clearly noisier and more biased than conventional occurrence data but they can be filtered by the confidence level of the taxonomic automatic classifier used in the app and they have the advantage of being produced in huge quantities.
2. Occurrences of other kingdoms: to investigate how knowledge of the presence of non-plants organisms can help predict the presence of plants species, we provided a large training set of occurrences from other kingdoms coming from the GBIF platform.
3. A better quality test set: to ensure the reliability of our evaluation, the occurrence data of the test set were restricted to expert data with the highest species identification certainty and high geographical accuracy (lower than 50 m). Last but not least, the test occurrences were sampled in order to avoid, as

[6] `https://www.gbif.org/`

much as possible, biases of spatial coverage and in the species representation. By this way, it contributes to give relatively more importance to rare species and scarce areas.

In the following sections, we describe in more details the data produced and the evaluation methodology used. We then present the results of the evaluation and the analysis of these results.

2 Dataset

2.1 Train occurrences

Pl@ntNet raw data. (`PL_complete`) This data is directly pulled from [4]. Pl@ntNet[7] is a smartphone app using machine learning to identify plant species from pictures submitted by a broad public of users. For each submission, also called a query, the Pl@ntNet algorithm answers a distribution of probability values across the targeted taxonomic referential. If the users allows it, the query's geolocation is also stored. In the provided training data, we used all accurately geolocated queries (with maximum 30 meters uncertainty) in France from the beginning of 2017 to the end of October 2018. Each geolocated occurrence is labelled with the species of higher identification probability. This dataset is thus very heterogeneous in species identification quality, due to the high variability of the image quality submitted by users. The confidence score is provided to GeoLifeCLEF participants as specific field in this dataset, who can use it to account for identification uncertainty in their models. This data set contains 2,377,610 occurrences covering 3,906 plant species.

Pl@ntNet filtered data. (`PL_filtered`) We proposed a filtered version of the previous dataset based on species identification quality. We only kept the occurrences for which the first species probability value was above 0.98. This score has been determined by expert to give a reasonable degree of identification confidence. This set of 237,087 occurrences covers 1,364 species.

GeoLifeClef 2018. (GBIF) Train and test occurrences datasets from the previous year edition [5] were merged to feed the current challenge. Those plants occurrences were extracted from the Global Biodiversity Information Facility [8]. This set of occurrences is around ten times smaller than the Pl@ntNet dataset, as shown in Figure 1. Within this dataset, occurrences are often aggregated on a same geographic point, which denotes uncertain or degraded geolocation. However, the geolocation certainty field is often missing. It contains $281,952$ occurrences covering $3,231$ plant species.

[7] https://plantnet.org
[8] `https://www.gbif.org/`

Occurrences of other kingdoms. (GBIF) This data source is made of species that are not plants, but may interact somehow with plants (e.g. trophic, pollination, symbiosis, use of plant as habitat or shelter), and are thus likely to carry interesting correlations with plant species presences. None of those species are in the list of species to predict in the test set (which are only plant species). Those occurrences have also been extracted from the GBIF; based on the following filters: { Basis of record: Human, Location : include coordinates, Country or area : France }. We extracted occurrences from 7 non-plant taxonomic groups:

- Chordata/ Aves (8,000,000).
- Chordata/ Mammalia (1,300,000)
- Chordata/ Amphibia (300,000)
- Chordata/ Reptilia (200,000)
- Arthropoda/ Insecta (3,250,000)
- Arthropoda/ Arachnida (70,000)
- Fungi/ Basidiomycota (50,000)

It contains 10,618,839 occurrences in total covering 23,893 taxa.

Taxonomic and geographic filters applied to all datasets. Because scientists do not name species by the same way in all regions of the world, many official lists of species names, called referentials, co-exist. There are no exact matching between them (in particular because of the new scientific knowledge acquired during the period between the creation of two separate lists) except those suggested by the scientific latin names themselves. In our case, the distinct data sources don't use the same referentials. Furthermore, distinct species names might be considered as redundant (synonyms) in some referentials. GBIF uses its own referential made from several taxonomic referentials, and GBIF occurrences may not be at the species taxonomic level, but at sub-species, or genus, etc. Pl@ntNet data includes occurrences from several plants taxonomic referentials (like The Plant List[9], GRIN[10], the French National plant list, etc.).
Thus, for attributing species identifiers in GeoLifeCLEF, it was important to first match all occurrences names to a single taxonomic referential adapted for the French Flora. We chose to use Taxref v12 [11] referential. We only kept names matching Taxref v12 according to an exact matching algorithm (R script provided on Github [12]). Some true species might have been lost due to distinct spelling between the GBIF taxonomy and Taxref.
We only kept points falling inside the French territory (Polygon from GADM[13]) or inside a 30 meters buffer zone, to account for geolocation uncertainty. Finally, occurrences were randomly shuffled to avoid any bias introduced by their order of use.

[9] `http://www.theplantlist.org/`
[10] `https://www.ars-grin.gov/`
[11] `https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en`
[12] `https://github.com/maximiliense/GLC19/blob/master/GITHUB_taxonomic_and_spatial_filtering.R`
[13] `https://gadm.org/`

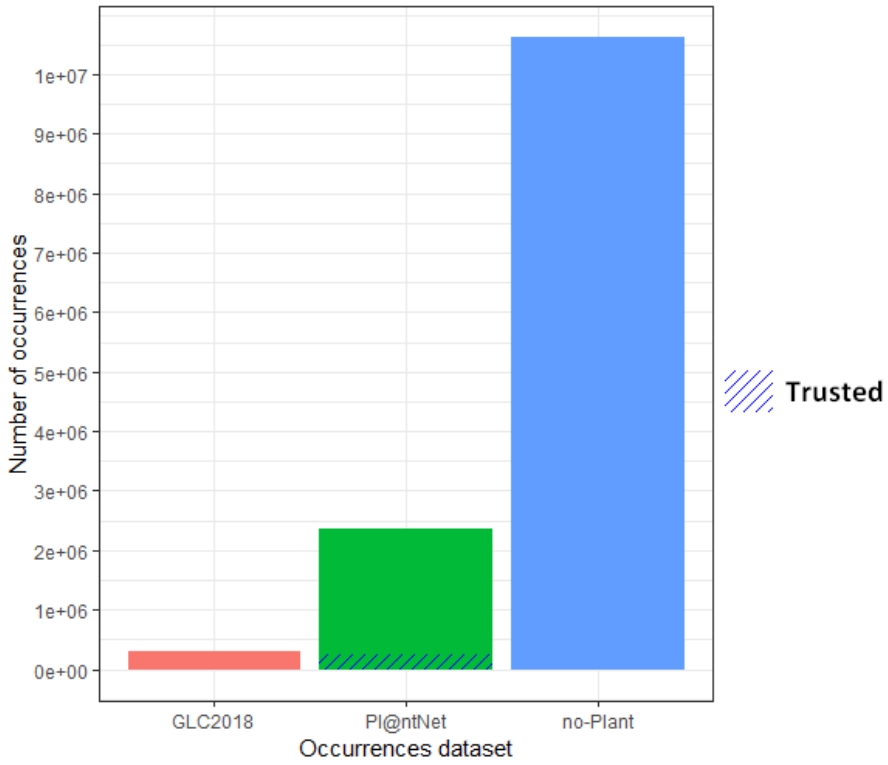

Fig. 1. Number of occurrences per training dataset. Trusted occurrences were determined from Pl@ntNet species identification engine certainty score.

2.2 Environmental data

Geographic rasters. The geographic and environmental data proposed to participants are a compilation of geographic rasters. The variables represented are often used for the purpose of species distribution modelling, especially for plants. The nature of values stored in the rasters are quantitative (bioclimatic, topological, hydrographical and evapo-transpiration variables), ordinal (pedological variables) or categorical (land cover). The rasters are extracted from the data repository of Botella [3], where readers can find a detailed description.

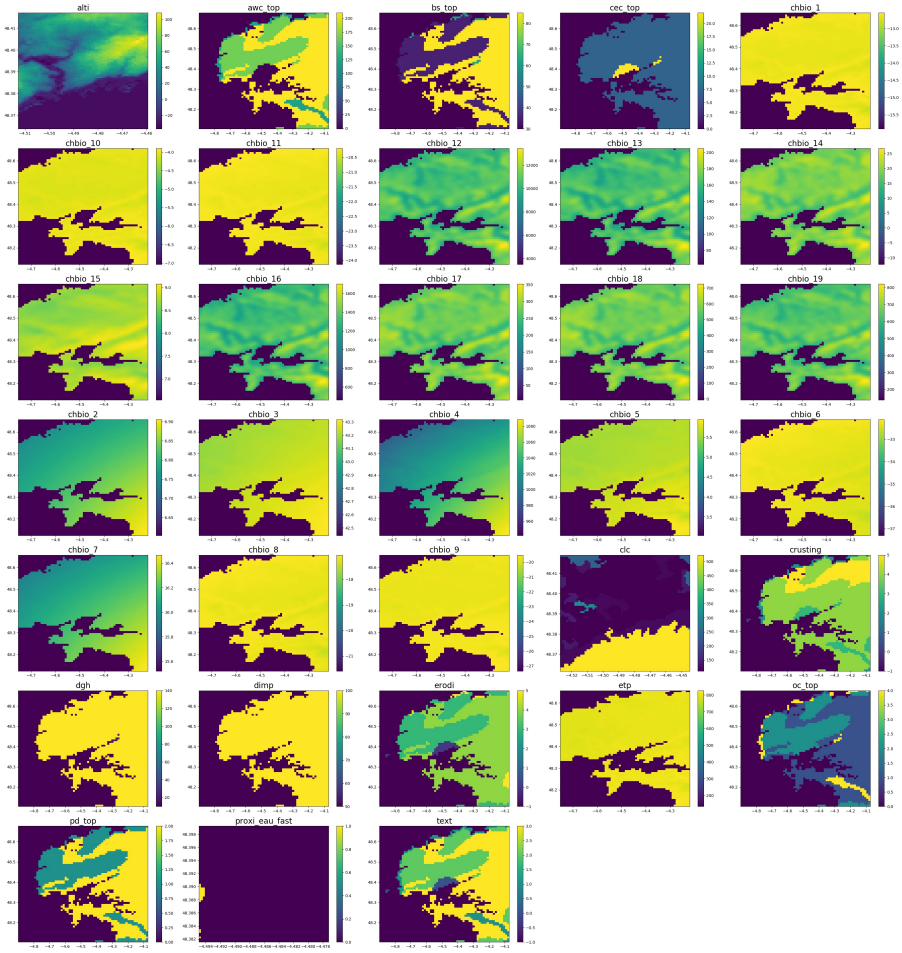

Fig. 2. Patch extracted at the city of Brest, France.

Tensors extraction. To facilitate the learning of representations taking into account the spatial structure of the environment, we provided a Python toolbox[14] allowing to extract local environmental tensors from any position in the rasters. By default, it extracts for each raster a $64x64$ pixels patch centered on the target position and aggregate the patches from all rasters in the form of a tensor of size nx64x64 where n is the number rasters.

2.3 Test data

We have chosen an independent and unpublished source dataset of occurrences for the test set. It is extracted from the *SILENE* database maintained by the *Conservatoire Botanique Mediterranéen* [15]. Those observations come from various providers including the conservatory himself, but also national parks, botanical associations or impact study consultants. We removed species (i) that were not present in the train set, (ii) vulnerable species according to the SINP referential *"espèces sensibles"* [16], (iii) and species that are at least *vulnerable* according to the IUCN red list [17]. This dataset has a high degree of identification certainty because only botanical experts contribute to it. Its geolocation certainty is under 50 meters. We used random weighted selection scheme to draw 25,000 test occurrences among the 700,000 of the initial set noted S. We compute, for each occurrence s_i in S a weight w_i:

$$w_i = 1/(n_i \times r_i)$$

Where r_i is the number of species in the neighborhood of s_i defined by a circle of radius d. n_i is the total number of occurrences in the neighborhood. We define the spatial scale $d = 2$ kilometers. With these weights and the following algorithm, we guaranty that (i) test occurrences are uniformly distributed in the geographic space at scale 2d, (ii) there is as many occurrences of each present species on neighborhoods of radius $2d$. We then draw the test occurrences from S without replacement, through the following algorithm:

- Initialize the bag of test occurrences $S' := S$ and the test set $T = \emptyset$.
- Randomly draw an occurrence in S', say i.
- Draw a scalar $z \sim U(0, max(w_1, ..., w_{|}S|))$.
- If $z < w_i$, remove i from S' and add it to T, otherwise leave it in S'.
- Stop if $|T| = 25000$, otherwise we go back to step (1).

3 Task description

For every occurrence of the test set, the evaluated systems must return a list of 50 species maximum, ranked without ex-aequo. The main evaluation metric

[14] https://github.com/maximiliense/GLC19
[15] `http://flore.silene.eu/index.php?cont=accueil`
[16] `http://www.naturefrance.fr/languedoc-roussillon/referentiel-des-donnees-sensibles`
[17] `https://uicn.fr/liste-rouge-flore/`

used is the top 30 accuracy (TOP30). We provide its expression hereafter:

$$\text{TOP30} : \frac{1}{Q}\sum_{q=1}^{Q} 1_{\text{rank}_q \leq 30}$$

where Q is the total number of query occurrences x_q in the test set and $rank_q$ is the rank of the correct species $y(x_q)$ in the ranked list of species predicted by the evaluated method for the occurrence x_q.
A secondary metric is the Mean Reciprocal Rank (MRR), a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries ordered by probability. The reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer. We provide its expression hereafter :

$$\text{MRR} : \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{\text{rank}_q}$$

The MRR was used as main metric during last year edition. We compute it this year, in order to enable comparisons between two campaigns.

4 Participants and methods

61 participants registered to the challenge through the online platform, among which 5 participants managed to submit runs in times. A total of 44 runs were submitted. All participants runs methods are characterized by their types of model architecture, the occurrences and input data they used in table 6. In the following paragraph, we describe in more details the methodology of each team.

LIRMM, Inria, Univ. Paul Valery, Univ. Montpellier, France, 4 runs, [10] : This team used a single deep convolutional neural network architecture derived in four models. All models take as input the default environmental tensors extracted by the provided python toolbox (see section 2.1), with a one-hot encoding transformation for each category of the land cover variables (`clc`), inducing 77 layers images in the input of the model. The chosen architecture was an Inception V3 ([13]). Models were trained as classifiers, using a softmax output and a cross-entropy loss (also known as multinomial logistic regression). Model of run 27006 was trained on all occurrences of `PL_complete` and `glc18` datasets, while models 27004 used `PL_complete` with identification score ≥ 0.7, and 27005 used `PL_complete` with identification score ≥ 0.98 (filtered dataset). Furthermore, runs 27004 and 27005 were only trained on a subset of the occurrences: a sample of around 30K occurrences was drawn according to the same selection procedure as for the test set. Thus, all those models predicted only plant species. On the contrary, model 27007 was trained on all occurrences datasets including `PL_complete`, `glc18` and also `noPlants`. This one was trained to predict plant species and many animal species.

SaraSi, EcoSols, UMR 1222 INRA - Montpellier SupAgro, France, 5 runs, [12] : This team used mainly two types of models: a convolutional neural network (CNN) based on the environmental tensors in the same spirit as LIRMM (27086, 27087, 27088) with a customized architecture, and a deep neural network using only a vector of co-occurrences of non-plants taxa as input (27089, 27082). The CNN model architecture separates the feature extraction depending on the type of variables that is deal with. Indeed, it apply distinct convolutional layers to the three categories of environmental patches (continuous, ordinal and categorical). The extracted features are concatenated and used as input in a series of fully-connected layers. A noticeable technique of "categories embedding" was used for the categorical and ordinal patches. It transforms the one-hot encoded patches in a lower number of continuous valued matrices. Also, they addressed the class imbalance of the training set by optimizing a weighted cross-entropy loss so that occurrences of more abundant species were less numerous. They trained this model on the *PL_complete* dataset (27086) and on a reduced version of this dataset to test set species (27088). the run 27087 was like 27086 but trained longer. For the other approach they implemented a customized version of the Continuous Bag of Words model [8]. The input is a set of identifiers of the non-plant "super-taxa" occurring in the neighborhood. An embedding vector associated to the set of "super-taxa" is learned. A "super-taxa" is an aggregation of many species assumed to share a same type of interaction with plants. They were determined through experts knowledge.

SSN_CSE, SSN College of Engineering of Chennai, and VIT University of Vellore, India, 12 run, [7] : This team tackles the challenge with classical machine learning techniques. They relied on three datasets : (i) spatial position of the occurrences only, (ii) spatial position and punctual environmental vector at the position of the occurrence, (iii) spatial position and vector of the average value of the environmental variables within a 16x16 pixels square centered on the occurrence. As a baseline, the authors first propose a probabilistic model where the probability of a species depends on its frequency in the whole training set (Const. prior). In addition, the authors relied on three categories of models. They first used random forest with spatial coordinates only as input (27102), and boosted trees (XGBoost: 26997, 26996, 27O13, 27012, 26988) and artificial neural network (27069, 27070, 27064, 27067) for using either spatial positions, environmental vectors or both. For one neural network, the authors split the features in 5 groups and trained a neural network per group for which predictions are then combined to form a single model.

Atodiresein, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania, 20 runs [2] : This team based their runs on standard machine learning algorithms: nearest neighbors (K-NN), random forests (Rand. For.), boosted trees (XGBoost) and deep neural networks (ANN). Those algorithms were applied to either the *PL_complete* or *PL_trusted* datasets. They used either the

spatial coordinates or the environmental punctual values of a selection of 29 environmental variables, or the concatenation of coordinates and variables. All combinations of algorithms, occurrences data and input data were evaluated on a validation set and the best of them were submitted. They also carried ensemble predictions from those models (runs 26969, 26970, 26958, 27062, 26960, 26971, 26961, 26964, 26968). A partial explanation of the low performances of their runs is that they only answered a short list of species (maximum 5) for each test occurrences, which lowers down performances a lot, especially for the top30 metric.

Lot_of_Lof, Inra, France, 3 runs, [9] : This team used occurrences density estimation based on log-linear spatial in-homogeneous Poisson point processes (PPP). They used a restricted set of environmental variables to model the distribution of occurrences based on expert knowledge: `etp, alti, chbio_5, chbio_12, awc_top, bs_top, slope` and aggregated `clc` in 5 land covers categories. They built their models with the 141 test species having the most occurrences in the *PL_trusted* dataset. Run 27124 is the standard PPP, while runs 27123 and 27063 apply different corrections for spatial sampling bias.

5 Results and discussion

The TOP30 and MRR evaluation scores achieved by all submitted runs are provided in Figures 3 and 4 (numerical values of the TOP30 are also replicated in the third column of Table 6). As a complementary analysis, Figure 5 displays the average TOP30 accuracy obtained for each species in the test set as a function of the number of occurrences of this species in the test set.

These results contributes to drive the following findings:

The occurrences of the other kingdoms significantly improve plants prediction. This can be observed from the comparison of run 27007 and run 27006 of the LIRMM team which are all things equal except the use of the occurrences of other kingdoms. The TOP30 increases from 0.136 to 0.177, which represents an improvement of 30%. The use of the occurrences of the other kingdoms is therefore the main cause of the best performances obtained by this team with regard to the SaraSi team. From the ecological point of view, this suggests that the biotic interactions (competition, predation, facilitation) between plant species and other biological groups play a very important role in determining the distribution of the species. From a deep learning point of view, it means that the convolutional neural network is able to transfer a consistent knowledge from the domain of the other kingdoms to the plant domain. An architecture that aim at predicting so many species through mutual neurons (as run 27007) might be a more efficient design for learning those relationships than using the co-occurrences as input data (as did runs 27089, 27082). It would be interesting to investigate this by comparing the latter strategy with a model taking both environmental patches and co-occurrences as input.

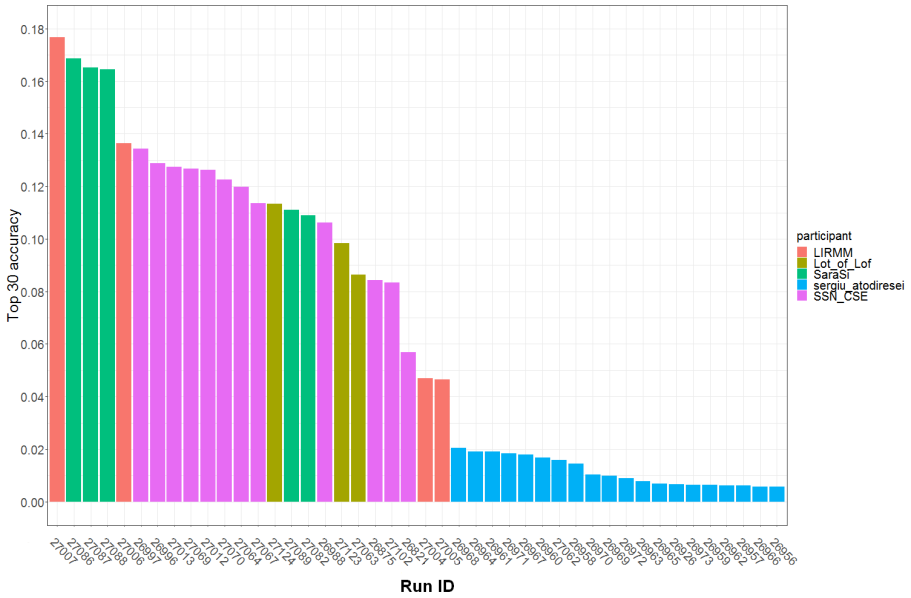

Fig. 3. Average Top30 accuracy per run and participant. It was computed over the 25,000 test occurrences. This was the official ranking metric for the task.

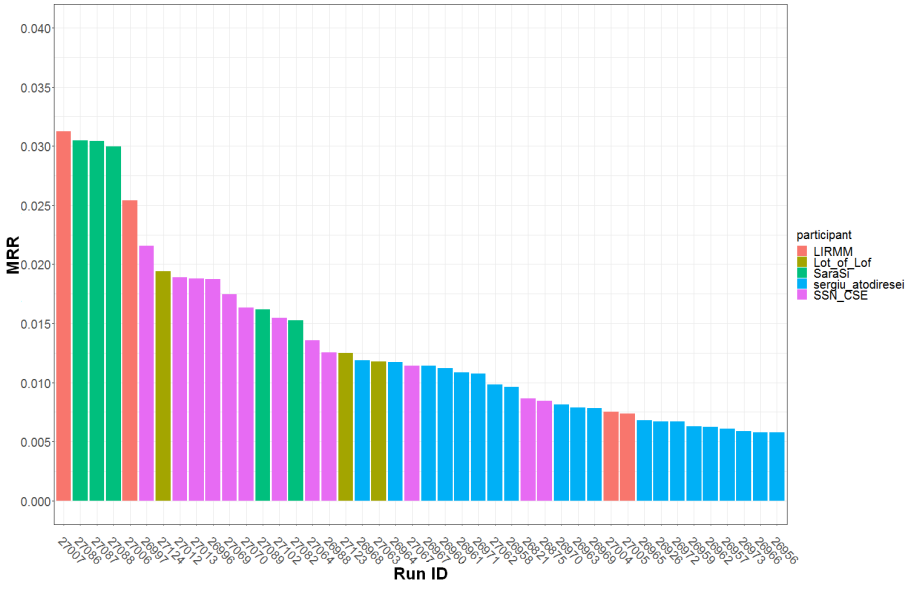

Fig. 4. Mean Reciprocal Rank per run and participant. It was computed over the 25,000 test occurrences.

Weighting the loss by species is better for predicting rare species. The CNN models learnt by the SaraSi team were based on a weighted cross-entropy loss penalizing the classes with more samples as a way to compensate class imbalance. Interestingly, it can be seen in Figure 5 that this significantly increased the ability of the model 27086 to predict the species having few occurrences compared to the winner CNN (run 27007) from LIRMM. Run 27086 is better than 27007 for more than 80% of the species. LIRMM team gave equal weights to all occurrences in the loss for training model 27007. It also shows how the most represented species hide the performances on the majority of species, which rarely occur. Giving more balanced weights across species is certainly important to achieve more robust predictions because the observation preferences across species vary a lot from one biodiversity dataset to another, as it is the case here between Pl@ntNet, the GBIF and SILENE.

The more complex the model, the better the prediction. The analysis of the column "model" of Table 6 suggests that, at least models using environmental inputs, can be ranked according to their performance as: (i) Convolutional Neural Network (CNN), (ii) Boosted trees (XGboost), (iii) Deep Neural Network (ANN), (iv) Poisson point processes, (v) K-Nearest Neighbors. This clearly shows a gradient from the models that integrate the most complex input data (CNN having the most complex with many channels of environmental images) and the most flexible architectures (CNN, XGBoost and ANN can fit very complex functions of their input data), to the models that are the most constrained by their input data (environmental vectors only) and with simple architectures (log-linear model of PPP, no optimized parameters for K-NN). This shows that the size of the available datasets and the complexity of the problem give a real advantage to complex statistical learning methods. More specifically, once again CNN results far exceeded those of the other methods which reinforces the results obtained in the last edition of the challenge. The CNN are likely to extract complex features of spatio-environmental patterns in their highest level neurons which are more suited to describe species habitats than environmental variables designed by experts. They may also captures spatial configurations of habitats that favor certain dispersion mechanisms, e.g. source-think colonization, or detect signatures of particular trophic assemblages.

The training of CNN can fail. Although the best models were based on CNNs, not all CNNs obtained so good results. Indeed, some runs based on CNNs were even worst than the prior ranking of species according to their global abundance (see $27004 \leq 26821$). Furthermore, non-submitted CNN models mentioned in a participant working note did perform less in validation than simpler approaches (see [7] 3.4). Model design (architecture, selection of environmental channels, management of categorical variables), regularization (optimization algorithm, use of dropout, learning rate and stopping rule policy), training data (especially size, see runs 27004 and 27005) and occurrence weighting scheme de-

termine jointly the implementation success.

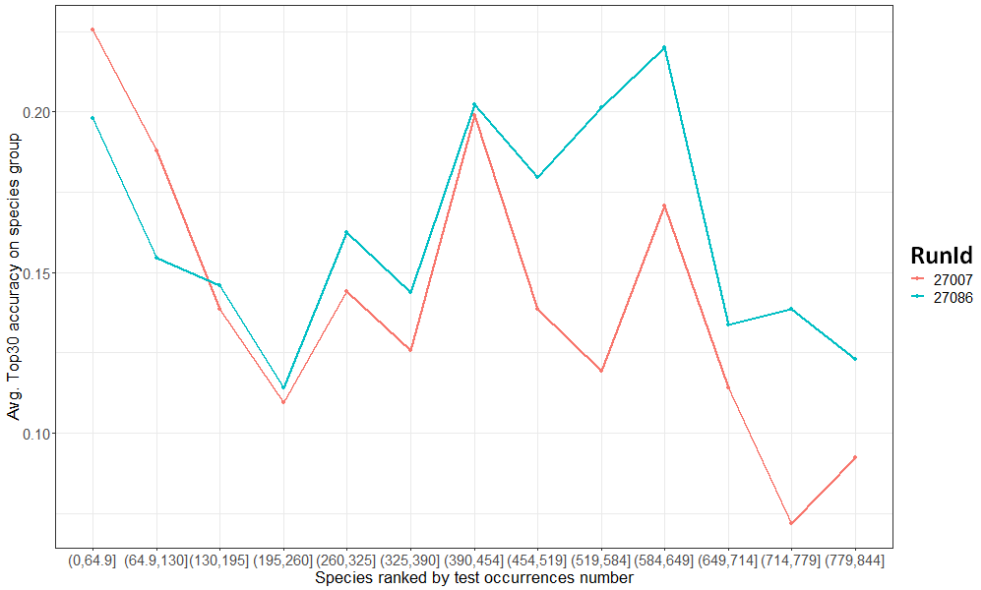

Fig. 5. Top30 accuracy averaged per species abundance class for the two best CNN models. Species were ranked by decreasing number of occurrences in the test set and then aggregated in 14 classes of abundances. For run 27086, each occurrence is weighted inversely proportional to the abundance of its species in the loss function.

Results of the MRR show that performances were globally lower than last year. Indeed, last year average MRR of the ten best runs was 0.039 while it is 0.024 this year. This large global performance gap is probably due to the difficulty of the test set, given that last year dataset was included in the training data. We note that the test set was not identically distributed, firstly because it was located on the Mediterranean region only, but also because the occurrences were sampled to avoid spatial and species biases. We know that all models predict less well rare species and under-sampled areas. Thus, this drop in overall performance supports the idea that the new test set has succeeded in giving greater importance to rare species and sub-sampled areas.
In absolute terms, the best run gives the good answer 20% of the times in its top-30. Thus, roughly speaking, even the best model gives generally a large majority of wrong species in its top-30 list. To give an order of comparison, the database Sophy [11] contains more than 35,000 exhaustive plant species inventories on plots generally not exceeding $400m^2$, and covers a wide range of environments in France. According to it, the species diversity in such plots is 25 in average

and rarely exceeds 70. There is thus large room for improvement in automated predictions.

6 Conclusion and perspectives

We now come back on the main outcomes of this task and discuss its perspectives.
LIRMM best CNN successfully integrated many non-plants species occurrences in their models predictions to better extract spatio-environmental patterns that more robustly predict plants species. It suggests that the global biotic assemblage highly determine the plant assemblage through underlying species interactions, and the multi-species prediction proved again to be a good deep learning strategy to account for it. This is the main new outcome of this year's edition. However, there should be significant room for improvement in the implementation of this approach. Indeed, LIRMM indicated that the winning model training couldn't be finished for time constraints reasons. Furthermore, light and customized models architectures accounting for the different variables natures seem more adapted to the problem than heavily parameterized state-of-the-art image classification architectures. Indeed, SaraSi customized CNN architecture has performed better than the related LIRMM Inception V3 CNN with the same output. Merging the strengths of both strategies promises good improvements in the future.
A rich source of information that remains unexploited for this task is the high resolution satellite images data. For example, today, 50 cm resolution satellite images are freely available for research all over the french territory through the National Institute of Geography (IGN) [18]. Including such images as input in the current models would inform them about very local land cover type and thus give much finer resolution prediction, if one can efficiently handle the size of this data.
The philosophy of the evaluation was to favor models that are more robust to biases in the training data, especially the imbalance of species representation and the heterogeneous spatial coverage, both consequences of the reporting process heterogeneity. We can say that it is a success concerning species imbalance representation. Indeed, SaraSi achieved remarkably stable performances even for rare species through a per class weighting scheme in the cost function. A next step would be to account for spatial sampling heterogeneity, as we have seen that all methods still struggle a lot with scarcely reported areas.
Regarding the evaluation process on this problem globally, we put an effort this year in the quality of the occurrences identification, and corrected for the species imbalance bias and heterogeneous spatial coverage (due to the reporting heterogeneity). Our new evaluation strategy was quite discriminant across the methods, and lowered globally the computed results. In absolute terms, we have also seen that even the best model tends to rank a lot of relevant species (i.e. probably absent from the surroundings) before the good one. The problem of spatial prediction of plant species lists is objectively far from being solved. Still,

[18] `https://geoservices.ign.fr/documentation/geoservices/`

with the new areas of improvements that the task results pointed out, we are optimistic about the future methodological advances on the problem of location based species prediction.

Rk	runId	top30	participant	model archi.	occurrences	covariates	supp. info.
1	27007	0.1769	LIRMM	CNN	all	enviro. tensors	–
2	27086	0.1687	SaraSi	CNN	complete	enviro. tensors	–
3	27087	0.1653	SaraSi	CNN	complete	enviro. tensors	27286 trained longer
4	27088	0.1646	SaraSi	CNN	complete $\cap$ test	enviro. tensors	–
5	27006	0.1364	LIRMM	CNN	complete + glc18	enviro. tensors	–
6	26997	0.1342	SSN_CSE	XGboost	filtered	enviro.+coord.	16 pixels avg.
7	26996	0.1288	SSN_CSE	XGboost	filtered	enviro.\ clc	–
8	27013	0.1273	SSN_CSE	XGboost	filtered	enviro.+coord.	max depth=3
9	27069	0.1268	SSN_CSE	ANN	filtered	enviro. selec.	16 pixels avg.
10	27012	0.1263	SSN_CSE	XGboost	filtered	enviro.+coord.	16 pixel. avg. max depth=3
11	27070	0.1227	SSN_CSE	ANN	filtered	enviro. selec.	–
12	27064	0.1198	SSN_CSE	$(\text{ANN})^5$	filtered	enviro.+coord.	$\neq$ covariates
13	27067	0.1135	SSN_CSE	ANN	filtered	enviro.+coord.	$\neq$ covariates 16 pixel. avg.
14	27124	0.1135	Lot_of_Lof	PPP	filtered	enviro. selec.	test sp. subset
15	27089	0.1110	SaraSi	NN	all plants	co-occurrences	large embed.
16	27082	0.1090	SaraSi	NN	all plants	co-occurrences	small embed.
17	26988	0.1063	SSN_CSE	XGBoost	filtered	coord.	–
18	27123	0.0984	Lot_of_Lof	PPP	filtered	enviro. selec.	test sp. subset
19	27063	0.0864	Lot_of_Lof	PPP	filtered	enviro. selec.	test sp. subset
20	26875	0.0844	SSN_CSE	ANN	filtered	coord.	–
21	27102	0.0834	SSN_CSE	Rand. For.	filtered	coord.	–
22	26821	0.0570	SSN_CSE	Const. prior	filtered	–	–
23	27004	0.0470	LIRMM	CNN	complete score$\geq$0.7	enviro. tensors	test-like sampling
24	27005	0.0465	LIRMM	CNN	filtered	”	”
25	26968	0.0205	atodiresei	$(\text{Rand.For.})^2$	filtered + complete	(enviro.\bio$_2$,text,clc) +coord.	–
26	26964	0.0191	atodiresei	$(\text{Rand.For.})^2$	filtered + complete	(enviro.\bio$_2$,text,clc)	–
27	26961	0.0190	atodiresei	$(\text{1-NN})^2$	filtered + complete	(enviro.\bio$_2$,text,clc) +coord.	–
28	26971	0.0184	atodiresei	1-NN $\times$ RF	filtered + complete	(enviro.\bio$_2$,text,clc) +coord.	–

Rk	runId	top30	participant	model archi.	occurrences	covariates	supp. info.
29	26967	0.0180	atodiresei	Rand. For.	filtered + complete	coord.	–
30	26960	0.0168	atodiresei	3-NN× 5-NN	filtered + complete	coord.	–
31	27062	0.0159	atodiresei	1-NN × RF × ANN × ANN × XGB	filtered	(enviro.\bio$_2$,text,clc)	–
32	26958	0.0146	atodiresei	3-NN× 5-NN	filtered + complete	(enviro.\bio$_2$,text,clc)	–
33	26970	0.0102	atodiresei	1-NN × RF	complete	(enviro.\bio$_2$,text,clc) +coord.	–
34	26969	0.0099	atodiresei	1-NN × RF	filtered	(enviro.\bio$_2$,text,clc) +coord.	
35	26972	0.0089	atodiresei	ANN	filtered + complete	(enviro.\bio$_2$,text,clc)	–
36	26963	0.0079	atodiresei	Rand. For.	complete	(enviro.\bio$_2$,text,clc)	–
39	26973	0.0064	atodiresei	XGBoost	filtered	(enviro.\bio$_2$,text,clc) +coord.	–
40	26959	0.0063	atodiresei	1-NN	complete	coord.	–
41	26962	0.0062	atodiresei	Rand. For.	filtered + complete	coord.	–
42	26957	0.0061	atodiresei	1-NN	complete	(enviro.\bio$_2$,text,clc)	–
43	26966	0.0058	atodiresei	Rand. For.	complete	coord.	–
44	26956	0.0058	atodiresei	1-NN	complete	(enviro.\bio$_2$,text,clc)	–

Table 1. Results and summarized methodology description of all runs submitted to GeoLifeCLEF 2019. Symbols and abbreviations: $A + B$ means that variables/data B was added to A. $A \backslash B$ means that variables/data B where removed from A. $complete \cap test$ means that only test species occurrences from the complete dataset were used. Products ($\times$) and exponent notations in column "model archi." decompose an ensemble methods with its different models. Occurrences: complete=$PL\_complete$,filtered=$PL\_filtered$, all plants=$PL\_complete + PL\_filtered + glc18$, all=$PL\_complete + PL\_filtered + glc18 + nonPlants$. Covariates in model input: "enviro. tensors"=environmental tensors with spatial neighborhood", "enviro."=punctual values of environmental variables, "coord."= spatial coordinates.

Bibliography

[1] Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.C., Joly, A.: Pl@ ntnet app in the era of deep learning. In: ICLR 2017-Workshop Track-5th International Conference on Learning Representations. pp. 1–6 (2017)

[2] Atodiresei, Costel-Sergiu, I.A.: Location-based species recommendation - geolifeclef 2019 challenge. proceedings of CLEF 2019 (2019)

[3] Botella, C.: A compilation of environmental geographic rasters for sdm covering france (version 1) [data set]. Zenodo (2019), http://doi.org/10.5281/zenodo.2635501

[4] Botella, C., Bonnet, P., Joly, A., Lombardo, J.C., Affouard, A.: Pl@ntnet queries 2017-2018 in france. Zenodo (2019), http://doi.org/10.5281/zenodo.2634137

[5] Botella, C., Bonnet, P., Munoz, F., Monestiez, P., Joly, A.: Overview of geolifeclef 2018: location-based species recommendation. In: CLEF 2018 (2018)

[6] Joly, A., Goëau, H., Botella, C., Kahl, S., Poupard, M., Servajean, M., Glotin, H., Bonnet, P., Vellinga, W.P., Planqué, R., Schlüter, J., Stöter, F.R., Müller, H.: Lifeclef 2019: Biodiversity identification and prediction challenges. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 275–282. Springer International Publishing, Cham (2019)

[7] Krishna, Nanda, K.P.K.R.M.P.A.C.J.S.: Species recommendation using machine learning - geolifeclef 2019. proceedings of CLEF 2019 (2019)

[8] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[9] Monestiez, Pascal, B.C.: Location-based species recommendation - geolifeclef 2019 challenge. proceedings of CLEF 2019 (2019)

[10] Negri, Mathilde, S.M.J.A.: Plant prediction from cnn model trained with other kingdom species (geolifeclef 2019: Lirmm team). proceedings of CLEF 2019 (2019)

[11] Ruffray, P., B.H.G.r.G.H.M.: "sophy", une banque de données phytosociologiques; son intérêt pour la conservation de la nature. Actes du colloque "Plantes sauvages et menacées de France: bilan et protection", Brest, 8-10 octobre 1987 pp. 129–150 (1989)

[12] Si-Moussi, Sara, G.E.H.M.D.T.T.W.: Species recommendation using environment and biotic associations. proceedings of CLEF 2019 (2019)

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)